

RAW VERSUS STANDARDIZED INTELLIGENCE TEST SCORES

The paper by Prof. J. M. Throne, published in the British Journal of Mental Subnormality in June 1972 has led to a rebuttal by Dr. Bialer. Dr. Bialer's argument and Prof. Throne's reply are of general interest and are published below.—The Editor.

—Is Throne's Position Valid ?

IRV BIALER

New York State Department of Mental Hygiene
Child Psychiatric Evaluation Research Unit, U.S.A.

Throne (1972) has come to the complementary conclusions that: (a) standardized intelligence test scores do not constitute valid baselines against which to assess the effectiveness of any treatment afforded retarded individuals, and (b) raw test scores—**from which the standardized scores are derived**—can serve as perfectly valid bases for such assessment. The major premises from which Throne (1972) draws his conclusions are: first, the circumstances prevailing at the time of testing differ for the standardization group and the retarded testee since the latter is being assessed subsequent to a given treatment, and "... any normal standardization sample . . . will typically have received no treatment prior to testing" (p. 37); and second, "raw test scores reflect testee performances regardless of the circumstances under which they occurred" (p. 38).

The **major fallacy** in Throne's position is his stand that the standardization sample represents an **untreated** sample with regard to the performances measured by the given test. In point of fact, implicit in any standardized test score (whether from test of academic achievement, intellectual level or what have you) is the **assumption that it reflects the average extent to which the standardization subjects have profited from given experiences up to the time they were tested.** For example, in deriving fourth-grade norms on an achievement test, it is assumed that the sample has had first, second and third-grade training in that academic area. In other words, it is a necessary condition that some form of treatment, training or teaching—formal or informal—has occurred before the standardization data were gathered. Otherwise, standardization samples would come to the testing situation with their minds in the shape of John Locke's "tabula rasa."

If a prime function of standardization is the establishment of **norms** which reflect average behaviours of given groups, a prime function of testing any mentally retarded subject with such an instrument (prior to instituting any treatment) is to determine the extent to which he deviates from normality in the given characteristic. Such deviation (where it is found to occur) may imply one of two conditions: (a) **S** has been exposed to the same kinds of prior experiences as the standardization group and he has not profited from such experience at least as much as has the average standardization subject; or (b) **S** has not had the experience which is necessary to respond appropriately.

Each of these conditions points to the need for a different follow-up in treatment or training or teaching. The question of whether or not a given retarded individual has been exposed to given experiences can only be ascertained by reference to his history.

Under condition (a) above, the implication that **S** cannot profit from the (educational, social, intellectual) experiences afforded the general population represented by the standardization sample indicates the need for programmes taking this into account. Such a condition usually does not call for pre-post testing with the standardized instrument to assess effectiveness of a given programme. However, under condition (b)—which denotes lack of training comparable to the normative sample—the interventive approach would ostensibly be designed to raise the retardate's

level of effective behaviour towards "normality." (This is congruent with the current notion of "normalization" with regard to management of the mentally retarded.) The only way in which we can then determine if the given programme has been thus effective is to compare the subject's post-treatment (test) behaviour with the norms.

It would follow that, having established the retarded S's level of functioning **relative to the norm**, predictions of extra-test performances of such (post-treatment) children can then be made as logically as for any other children for whom the tests are used in a predictive fashion.

To use Throne's diamond Luminosity Quotient (LQ) analogy (1972, p. 37), the rough diamond (read "retardate") is compared to the LQ norms of polished diamonds of various sizes, shapes, etc. (read "standardization sample") before polishing (read "treatment") to determine how far below the norm it is, and then again after polishing to see if it has the normative value needed for saleability (read "effective behaviour"). In that case, treatment allows one to predict saleability by comparing the finished product with one **known to be saleable**. Comparisons of untreated to treated gems is valid if one needs to determine how much treatment might be necessary to render the former saleable. Thus, post-treatment intra-test testee performance can be expected to predict to extra-test testee performance to the extent that we wish to know how such extra-test performance compares to the norm (i.e., how **normal** has the given treatment, training or education succeeded in rendering the retarded subject?).

The **second fallacy** in Throne's (1972) position is his insistence that raw scores are always valid baselines for assessing treatment effectiveness.

First of all, in comparing the validity of raw and standardized test scores with regard to the circumstances under which they are obtained, Throne is apparently subliminally aware of, but chooses to neglect, the fact that raw scores are derived from the responses to test items of the standardization sample. Indeed, the standard score is often a function of the **average raw scores** at different levels of the standardization sample. To that extent, **the raw score is as much a standardized score as is the standard score**.

As noted above, for educational and/or other treatment/planning purposes, it is often necessary to know the extent of the child's deviation from the norm—or his relative standing in a given group. These variables cannot be assessed from raw scores alone. In addition, while a pre-post test comparison of a given child's **raw scores** will indicate if an increment has occurred as a function of a given treatment, neither the absolute magnitude of the increment—nor of the post-test score itself—can tell us if the interventive programme has succeeded in rendering the mentally retarded individual more like the normal one. This can only be done by comparing the deviant child's **standard score** with the norm, i.e., by comparing his post-test performance with the typical performance of the average child his age.

Throne (1972, pp. 39-40) also argues that, as opposed to standard scores, raw test scores are the only valid measures for assessing treatment effectiveness if one wishes to measure performance level relative to S himself. In a discussion of the clinical utility of the Wechsler Intelligence Scale for Children (WISC) with mentally retarded children, the present writer has observed elsewhere that:

Since the magnitude of a given raw score does not adequately reflect an individual's **relative standing** on a given attribute—with respect either to a given norm or to his own overall performance—raw scores are not directly comparable across attributes. However, standard scores express an individual's position relative to the criterion group in terms of units (such as standard deviations) which are directly comparable. Consequently, standard scores for sub-scale or subtest

portion's of (various) intelligence tests . . . provide a means of determining not only the individual's standing relative to the normative sample. . . . Such scores also allow for analyzing a given individual's strengths and weaknesses—in terms of the psychological and/or educational processes tapped by given subtests. . . . (Bialer, 1971, p. 6).

Since raw scores in general simply reflect the number and/or magnitude of correct responses to specific items, they bear no direct, observable relationship to each other which can tell us at a glance which subtest or achievement score is "better" or "worse" than any other. However, the standard score, which takes into account the average raw score on a given subtest or achievement test for a specific age or grade level, immediately tells us to what extent a given raw score deviates from the norm. This helps establish relative intra-individual strengths and weaknesses, either with regard to certain psychological processes or academic subjects, and points to needed areas of specific intervention.

Further elaboration of the issues is beyond the scope of the present discussion. However, in spite of Throne's alarm regarding the questionable scientific and/or ethical use of standardized test scores, the conclusion seems inescapable that standard scores, derived from standardization data, provide a valid basis for both inter- and intra-subject assessment with mentally retarded children, whether the assessment be for clinical or research purposes.

References

- Bialer, I. The valid utilization of standardized intelligence tests with the mentally retarded. Paper presented at the meeting of the American Psychological Association, Washington, D.C., September 1971.
- Throne, J. (1972). Raw versus standardized intelligence test scores as baselines for assessing effectiveness of treatment: Implications for the mentally retarded. *The British Journal of Mental Subnormality*, 18(1), 36-43.

II—Reply to I. Bialer

JOHN M. THRONE
University of Kansas, U.S.A.

The first and major fallacy Bialer attributes to me is my statement that the standardization sample represents an untreated sample with regard to the performances measured by the test. Bialer maintains that, quite to the contrary, the standardization samples have indeed received treatment; they have been subjected to "life's experiences." Of course they have, but these are of an entirely different order of discourse from those implicit in the concept of treatment. Life's experiences most assuredly do not impinge upon any standardization sample with anything approaching the form or intensity that is implied by the concept of treatment. Bialer's fourth-graders, assumed to have experienced grades one through three, are also assumed to have done so under circumstances having nothing to do with training or teaching in the sense of treatment—i.e., institutionalization, special education, private tutoring, etc. Neither have the samples of standardized achievement tests; therefore, subjecting any fourth-graders to standardized achievement testing is scientifically and ethically justified only if their scores—like those of the samples against which theirs must be matched in order to be evaluated—have been, are, and are expected to remain unconfounded by treatment—a caveat of course precluding their employment for baseline purposes, i.e., to evaluate treatment effectiveness.

We are talking here about **standardized** test scores—measures of performances on test items obtained under the same conditions prevailing for a standardization sample before, during and after the time they are examined on the same test items. Before and during testing, standardized conditions insure that the reflected intratest performances on test items by sample and subjects are comparable (internal validity).

After testing, they insure that their performances indicate comparable levels of extra-test performance on tasks bearing a functional relationship to the test items—that is, predictability (external validity). In and of themselves, the conditions under which the performance levels of samples on standardized tests are measured do not matter; they must be standardized only to insure that the performance levels of subjects will be measured under the same conditions as those of the samples. If they are not, so-called “standardized test scores” on subjects are not that at all; insofar as can be told, they are devoid of validity in both the internal and external senses of that term. In short, whether they are quotients of intelligence on human beings or of luminosity on diamonds, to be internally or externally valid, or both, standardized test scores must be obtained under standardized conditions of testing. Otherwise they are false scores, and no amount of wishful (and misleading) thinking that they indicate the amount of treatment needed to make either a dull intelligence or a dull diamond brighter, will make them true.

My second fallacy, according to Bialer, is my advocacy of raw test scores on standardized tests in lieu of standardized test scores. Bialer believes I am subliminally aware of but choose to ignore the fact that the latter are derived from the former. I must stoutly demur. My awareness is both supraliminal and attentive, so much so that I urge that the performances of subjects reflected in raw test scores on (for example) standardized intelligence tests but obtained under **non**-standardized conditions be employed as nuclei around which the development of intelligence, of course construed in purely behavioural terms (i.e., as represented by those performances), may be attacked through the same programming techniques that may be used in developing any behaviour. In which case, the raw test scores reflecting those performances are no less indices of intelligence for having been obtained under nonstandardized conditions; such scores are then unusable for comparative and predictive purposes, but this is irrelevant if they are intended to reflect the performances of subjects prior and subsequent to treatment aimed at improving those performances (as is implicit in the very baseline concept). Production of improvement renders its mere comparison and prediction via standardized testing academic, of course.

With the term nuclei I mean to imply the indicative, not exhaustive, nature of those performances reflected in the items on standardized tests. For example, on the Wechsler Adult Intelligence Scale, there are eleven subtests containing dozens of items reflecting performances which, even if scored for subjects under non-standardized conditions, are nonetheless representative of **an** aggregate of performances that have been accepted through usage by psychologists as denoting intelligence. Extrapolation from those WAIS subtests and items, or from those of any other standardized intelligence test, appears to me to be as logical a curriculum strategy to follow as any available to the scientist, clinician, or educator who would attempt to improve the intelligence of mentally retarded or other subjects. That is, those WAIS subtests and items, through professional consensus, have earned (if you will) the status of **intelligence indices** on whose bases programmes may be erected for developing intelligence in subjects for whom treatment along such lines is prescribed. For individual subjects (whether they are mentally retarded, average, or advanced), this surely is the prime, indeed the only, consideration; not how well (or poorly) they perform in relation to the average subjects in a population, as represented by the standardization sample of any standardized test, whether the WAIS or some other.

For a slightly revised copy of the original paper which is the subject of this exchange between Bialer and myself, readers may write: J. M. Throne, Centre for Mental Retardation and Human Development, 223 New Haworth, University of Kansas, Lawrence, Kansas 66045, U.S.A.